De-Identification of health records using *Anonym*: Effectiveness and robustness across datasets

Guido Zuccon[a,b], Daniel Kotzur[a], Anthony Nguyen[a], Anton Bergheim[c]

[a] *The Australian e-Health Research Centre (Commonwealth Scientific and Industrial Research Organisation), Level 5 - UQ Health Sciences Building 901/16, Royal Brisbane and Women's Hospital, Herston, QLD 4029 Australia*

[b] *School of Information Systems, Queensland University of Technology, Y Block Level 6, Gardens Point Campus, Brisbane, QLD, Australia*

[c] *Cancer Institute NSW, Australian Technology Park Level 9, 8 Central Avenue, Eveleigh NSW 2015, Australia*

Abstract

Objective: Evaluate the effectiveness and robustness of Anonym, a tool for de-identifying free-text health records based on conditional random fields classifiers informed by linguistic and lexical features, as well as features extracted by pattern matching techniques. De-identification of personal health information in electronic health records is essential for the sharing and secondary usage of clinical data. De-identification tools that adapt to different sources of clinical data are attractive as they would require minimal intervention to guarantee high effectiveness.

Methods and Materials: The effectiveness and robustness of Anonym are evaluated across multiple datasets, including the widely adopted Integrating Biology and the Bedside (i2b2) dataset, used for evaluation in a de-identification challenge. The datasets used here vary in type of health records, source of data, and their quality, with one of the datasets containing optical character recognition errors.

Results: Anonym identifies and removes up to 96.6% of personal health identifiers (recall) with a precision of up to 98.2% on the i2b2 dataset, outperforming the best system proposed in the i2b2 challenge. The effectiveness of Anonym across datasets is found to depend on the amount of information available for training.

Conclusion: Findings show that Anonym compares to the best approach from the 2006 i2b2 shared task. It is easy to retrain Anonym with new datasets; if retrained, the system is robust to variations of training size, data type and quality in presence of sufficient training data.

Keywords: Conditional Random Fields, Pattern Matching,
De-identification, Health records.

1. Background

Electronic health records (EHRs) often contain personal health information (PHI) that can uniquely identify a patient. The United States's Health Information Portability and Accountability Act (HIPAA) has stipulated 17 categories of PHIs that must be de-identified, the most prevalent are outlined in Table 1.

Access to EHRs outside of the primary health provider and the sharing of such data for research purposes is fundamental for critical data mining and information retrieval tasks in the health domain; for example, the identification of adverse drug reactions or patient recruitment for clinical studies [1, 2]. However, PHIs are pervasive in unstructured portions of EHRs, which undermines access and sharing of such important data [3].

De-identification is the process of removing PHIs from medical records.

Table 1: Subset of the United States's Health Information Portability and Accountability Act personal health identifiers types considered for the evaluation of Anonym.

PHI Type	Meaning
Patients	First, middle and last names of patients and their family members (including initials of names).
Doctors	Similar to patients category, includes names and initials of health professionals.
Dates	All numerical and literal reference to dates, including years and days of the week.
Hospitals	Names of medical facilities and practices.
IDs	Any combination of digits and letters that refer to medical records, patient numbers, accession numbers, doctors identifiers, laboratory identifiers, etc.
Locations	Names of cities, regions and states, as well as addresses, zip codes and building names.
Phone numbers	Any reference to landline, fax and mobile phone numbers or phone extension numbers.

Manual de-identification of electronic health records is time and resource consuming. Dorr et al. [4] found that on average 87.2 ± 61 seconds are required to manually de-identify a narrative text of an EHR; an EHR on average contains 7.9 ± 6.1 PHI entities.

Anonym is a software tool developed at the Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), that automatically de-identifies EHRs. Anonym is based on the combination of conditional random fields (CRF) classifiers, informed by a number of linguistic and lexical features and pattern matching techniques. The de-identification method used in Anonym is described in Section 3. The results of the empirical evaluation reported in Section 5 shall show that, if enough training data is provided, Anonym is capable of effectively de-identify free-text EHRs, irrespective of type, source or quality of data. In addition, results demonstrate that Anonym is comparable to the best state-of-the-art de-identification system proposed by Uzuner et al. [5]. In addition, the results also demonstrate that retraining is necessary when changing datasets.

2. Related work

Two areas of related work are reviewed: de-identification and named entity recognition.

2.1. De-identification

Research on de-identification of EHRs has flourished as a result of the introduction of the 2006 Integrating Biology and the Bedside (i2b2) dataset shared task [5]. This shared task provided an evaluation framework for de-identification, consisting of a dataset of manually annotated medical discharge summaries populated with ambiguous PHIs and metrics to measure the performance of de-identification systems. Uzuner et al. [5] provide an overview of systems that participated in i2b2. The techniques used by participants included conditional random fields, rule-based approaches, hidden Markov models (HMM), and support vector machines. The best system in i2b2 was developed by Wellner et al. [6]. Their system is similar to the approach considered in this paper: both use CRFs to label tokens and regular expressions to form one of the feature classes. However, our approach differs in that we do not use lexicons of locations and English words and we consider additional features such as part of speech.

Uzuner et al. [7] have studied the role of local context (i.e. the words that are immediate neighbours of the target PHI or that have immediate syntactic relation with it) for de-identification when using support vector machine classifiers. They observed that features that thoroughly capture local context are beneficial to the PHI de-identification task. While not relying on local context features as thoroughly as Uzuner et al. [7], Anonym does use features that implicitly capture local context information, such as token n-grams and part-of-speech.

An overview of approaches to PHI de-identification is provided by Meystre et al. [8]. From their analysis, they concluded that methods based on linguistic resources, such as dictionaries, tend to perform better with rarely mentioned PHIs. Vice versa, they found that machine learning techniques better generalise to PHIs that are not mentioned in dictionaries, although machine learning tends to have problems identifying PHI types that rarely occur in the training corpus. Rule-based techniques and machine learning algorithms have been recently integrated in the stepwise hybrid approach proposed by Ferrandez et al. [9]. Anonym uses rule-based techniques in its pattern-matching component for feature generation.

Recent work has focused on semi-supervised or iterative approaches that improve the human-supervised de-identification workflow process as a whole, rather than producing a fully automatic de-identification system. Hanauer et al. [10] constructed statistical de-identification models by iteratively performing (i) annotation of a small EHRs sample; (ii) training of a CRF model; (iii) automatic identification of PHIs on a small sample of unseen data; (iv) manual correction of the errors on the unseen data; and (v) retraining of the model. Boström and Dalianis [11] used active learning to train a random forest classifier to detect PHIs from Swedish EHRs. They also investigated different strategies to select the most discriminative samples for online manual annotation.

In a previous paper [12], we presented the approach underlying Anonym and initial results that showed our tool is comparable to state-of-the-art approaches on the 2006 i2b2 shared task. In that work, we have also briefly investigated the effectiveness on a small set of pathology reports supplied by an Australian cancer registry. This article extends that work by considering (1) additional datasets, including a larger set of cytology and pathology reports from a statewide Australian cancer registry and 1,885 clinical notes from the MTSamples dataset [13]; (2) further investigation of the adaptability of Anonym across the different datasets.

2.2. Named entity recognition and conditional random fields

De-identification is a specialisation of named entity recognition (NER), i.e., the task of recognising references in text to information units like names (e.g., persons, organisations, locations) and numeric expressions (e.g., dates, money). While early NER systems were based on highly engineered rules, the most recent and successful approaches adopt supervised machine learning to automatically induce recognition rules from a corpus of training examples. Popular supervised algorithms for NER include HMMs, decision trees, maximum entropy, support vector machines and CRF. A survey of NER models, common features, and evaluation techniques is given by Nadeau and Sekine [14].

Anonym is based on the conditional random fields approach to learn PHIs and then identify new occurrences of PHIs from unseen data. A CRF is a discriminative undirected probabilistic graphical model that, given an observed sequence, defines a log-linear distribution over labelled sequences [15]. Mathematically, given an observed sequence x, a CRF predicts a label y from the set of possible labels Y if y maximises the conditional probability $p(y|x)$, i.e., if $p(y|x)$ is greater than any $p(y^*|x)$, for all y^* in $Y \setminus \{y\}$. This conditional nature of CRF is the key characteristic distinguishing CRF from HMM; it also means that the independence assumption necessary to ensure tractable inference in HMM is relaxed in the CRF approach.

The CRF approach underneath Anonym uses, among others, features generated by a set of pattern matching rules (regular expressions). This feature generation approach is similar to that of Collins [16], who introduced pattern features that map tokens onto a set of patterns.

3. Anonym: de-identifying EHRs with CRF and pattern matching

Anonym consists of three main modules: (i) the automatic feature generation component, (ii) the model training component that uses the features generated by the first module, and (iii) the classification component which applies the learnt model to unseen data. A fourth module is responsible for the generation of PHI surrogates consistent with those identified and the replacement of the identified PHI with its surrogate. This component has not been used to post-process the PHIs identified by Anonym in this work. Instead, we used this component to pre-process the data of two of the three datasets considered here as they could not be distributed with the original text representing the identified PHIs. The component is briefly described in Section 4.2. Next, we describe the feature generation component of Anonym. We do not describe the other two components in Anonym as they resemble standard CRF classifiers. Note, however, that the training module allows for selecting which features are used for learning.

Anonym is able to extract a number of lexical and linguistic features, grouped in seven general families: (1) basic features, which comprise word shapes (e.g. the presence of capitalised characters at the beginning of the word or across the whole word) and character n-grams ($n = 6$); (2) disjunctive features, which capture disjunctions of words and word shapes within windows of words; (3) short character n-grams (i.e. 3-grams) in place of the 6-grams used as basic features; (4) combination of short words, which creates a feature combining adjacent words of length three or less; (5) position features, which capture the position of a word in the sentence and in the PHIs. In addition we separately extracted features using (6) part-of-speech obtained from the Stanford Part of Speech Tagger [17], and (7) pattern matching techniques, i.e., by defining a set of regular expressions and assigning specific labels to tokens that match these regular expressions.

While lexical and linguistic features, such as word shapes and part-of-speech, are commonly used for de-identification, the extraction of an additional feature set using pattern matching techniques via regular expressions is a key characteristic of Anonym. In the current implementation, Anonym identifies patterns for the following categories: `Date`, `DateLabel`, `Time`, `Phone` and `PhoneAreaCode` (with different patterns for United States and Australian numbers), `PhoneLabel`, `NameLabel`, `Numeric`, `PersonTitle`. The category `Numeric` includes patterns that match occurrences of digits, i.e., that match the regular expression `"([0-9]1,)"`. The category `PersonTitle` refers to occurrences of name references; regular expressions for this category identify tokens matching a list of name titles (e.g., Dr., Prof.) and capture multiple references to names. Example regular expressions for the other categories are given in Table 2. A feature value is assigned to a token if matches one of the patterns, e.g. if a string is matched by a regular expression identifying a possible date, the value `DATE` is assigned to the pattern matching feature. These regular expressions were formed by analysing the training set of the i2b2 dataset [5]; in addition, relevant patterns were adapted to mimic Australian conventions for expressing dates, phones, etc.

4. Evaluating Anonym

An objective of this paper is to evaluate the effectiveness and robustness of Anonym. To do so, we first investigate whether Anonym consistently achieves

Table 2: Examples of regular expressions used by Anonym to identify pattern matching features for different categories. Regular expressions are reported as the strings used as targets for regular expression matching in Java.

<table>
<tr><th>Category</th><th>Pattern</th></tr>
<tr><td>PersonTitle:</td><td>(name|surname(\\s?)([:-]?))</td></tr>
<tr><td>DateLabel:</td><td>((date|(d(.?)d(.?)b(.?)))([:-]?))</td></tr>
<tr><td>PhoneLabel:</td><td>(((phone)|(phone number)|(telephone)|(mobile)))(\\s?)([:-]?)</td></tr>
<tr><td>Phone (AU):</td><td>"((([0-9]{4})([\\-\\.])([0-9]{4}))|([0-9]{2}([\\- \\.])[0-9]{3}([\\- \\.])[0-9]{3})|([0-9]{8}))"</td></tr>
<tr><td>PhoneArearCode (AU):</td><td>((([+]?([0-9]{2})([\\. \\-]?)[0-9])|((\\()(\\b)[0-9]{2}(\\)))))|((\\b)))</td></tr>
<tr><td>Time:</td><td>(\\b)(([0]?[0-9]|[1][1-2])(([\\.\\- \\:])[0-5][0-9]){1,2})([\\. \\- \\:]?)(am|pm)</td></tr>
<tr><td>Date:</td><td>(?<!([\\\\\\/ \\-]([0-9]{1,})?))((([01]?[0-9]|[2][1-4])(([\\. \\- \\:])[0-5][0-9]){1,2})(?!(([\\. \\- \\:][0-5][0-9])?)([\\. \\- \\:]?(am|pm))))</td></tr>
</table>

high effectiveness across the considered datasets (RQ1). The comparison of results across datasets will allow us to assess whether different training size and class distributions (of PHIs) affect effectiveness and in particular if high effectiveness is associated with large training data and low diversity across instances of PHI types (RQ2). This analysis will also highlight whether there is a unique combination of features that provides the highest effectiveness across all datasets (RQ3). The last two aspects contribute to evaluate the robustness of our software. To complement this analysis, we also study whether approaches trained on one dataset adapt to other datasets (RQ4).

4.1. Evaluation datasets

Three datasets were used to evaluate Anonym. The 2006 i2b2 shared task dataset consists of 889 medical discharge summaries annotated for evaluating PHI de-identification approaches (of these, 669 documents are commonly used for training, while the remaining 220 are used for testing). Details about this dataset are provided by Uzuner et al. [5]. This dataset is used to compare Anonym with state-of-the-art approaches studied in the relevant literature.

A second dataset was formed using 1,885 clinical notes from the MTSamples corpus. The dataset was manually annotated by reviewers at the University of California at San Diego, following the procedure outlined by South et al. [13]. The annotations in this corpus refer to the broader set of PHIs defined by HIPAA, including clinical eponyms, health care units, organisation names, etc. As detailed later, we only used a subset of these PHI types. PHIs identified by reviewers were automatically replaced with realistic surrogates produced by the Anonym module outlined in Section 4.2.

A third dataset was compiled using pathology and cytology reports obtained from Cancer Institute New South Wales[1]; we refer to this dataset as *CINSW*. The dataset contains 852 free-text reports acquired from paper source using an optical character recognition (OCR) software. The CINSW dataset used here is about four time larger than that used in the prelimi-

[1] With ethical approval granted by the New South Wales Population & Health Services Research Ethics Committee.

nary experiments reported in [12]. Reports in CINSW differ from those in the i2b2 and MTSamples datasets because of the linguistic and orthographic conventions in Australia vs. the U.S.. In addition, these documents contain OCR errors and loss in formatting, which may cause lower effectiveness from automatic de-identification tools. Details of OCR errors found in the data obtained from Cancer Institute New South Wales are reported in [18]. Manual annotation of PHIs in the CINSW dataset was performed by two authors of this paper. The process was aided by a graphical interface that highlighted patterns identified by the regular expressions described in Section 3. Each report was first annotated by one author and then manually reviewed by the second author. Manually identified PHIs were replaced using Anonym, similarly to the previous datasets. Note that patient names were not present in the PHIs identified in this dataset. This is because in the reports acquired from Cancer Institute New South Wales, names of patients are only present in the headers of the reports, which were excluded by the template used in the OCR process.

Not all the 18 PHI types identified by HIPAA are present across all three datasets. We restrict our experiments only to the subset of PHIs that is most commonly present across all datasets; considered PHIs are outlined in Table 3 along with occurrence statistics. Note that in the i2b2 dataset names of patients and doctors are annotated as separate PHIs; in the MTSamples dataset, there are different annotation types depending on whether a name refers to a patient, a relative of the patient, or another person, and there is

Table 3: Distribution of instances across personal health identifier types in the datasets considered by the evaluation of Anonym. For the i2b2 dataset, statistics are collected across both training and testing data.

	# of samples		
PHI Type	i2b2	MTSamples	CINSW
DATE	6,816	1,667	975
PATIENT	929	-	-
DOCTOR	3,386	-	-
NAME	-	286	1,725
AGE	16	1,748	13
ID	4,763	95	747
HOSPITAL	2,305	-	-
LOCATION	243	-	-
INSTITUTION	-	1,170	177
PHONE	222	7	540

no type that explicitly indicates names of doctors[2]. In the CINSW dataset there is no mention of names of patients, as detailed previously. We conflate references to names into the PHI type NAME in the MTSamples and CINSW datasets, while we keep references to patients and doctors separated in the i2b2 datasets to allow for direct comparison with results presented in the literature. Similarly, we distinguish between mentions of locations and hospitals in i2b2, while conflate relevant mentions into the type INSTITUTION for the MTSamples and CINSW datasets.

[2] These are instead grouped in the annotation type `HealthCareProviderName`.

4.2. Generation of PHIs surrogates

A module within Anonym has been implemented to automate the replacement of PHIs with realistic surrogates. Person names, institutions and locations are replaced with surrogates from a candidate list. Anonym allows a list of candidate surrogates to be provided by the user; alternatively, if no list is provided, or the corresponding setting is enabled, candidate surrogates are scrapped from the Web, using resources such as the White Pages website[3] and relevant Wikipedia pages[4]. To maintain consistency with the original PHIs, the tool attempts to identify the format used by the PHI, e.g., for a person's name, if it is in the format *FirstName MiddleName SecondName* or *FirstName M. SecondName*, etc. The surrogate string used to replace the original PHI is then matched to the correct format. If no format is automatically identified by Anonym, then the string is formatted according to a randomly selected known format. In the experiments reported in this article, we used automatic generated PHIs scrapped from the Web.

Dates are instead automatically shifted (forwards or backwards[5]) by a random offset. This offset is generated for each document and is applied to

[3] `http://www.whitepages.com/`; for example `http://www.whitepages.com/ind/a` is used to gather names of people whose surname starts with the letter A. URLs were last accessed on March 17, 2014.

[4] For example, hospitals names are gathered by mining the webpage at `http://en.wikipedia.org/wiki/List_of_hospitals_in_the_United_States/` (Accessed: March 17, 2014).

[5] An allowed date range is used to maintain dates constrained within a time period to avoid the generation of unrealistic data surrogates, i.e., those in the future or in the distant past.

all dates in that document; thus dates in one document may all be shifted backwards 2 months, while dates in a second document may be shifted forwards 10 days.

Phone numbers are randomly generated using the same number of digits used in the original PHIs. We implemented different phone number generators tailored to United States and Australia, with restrictions on the area code used. However, we did not use these restrictions in the experiments reported in this article.

ID numbers, accession numbers, and other relevant codes are replaced with randomly generated codes that follow the same structure of the original PHIs, i.e., by replacing a random digit with each digit in the original PHI, and a random character for each character in the original PHI.

4.3. Experimental settings

The features described in Section 3 were used to build the CRF classifier. The model was trained using features extracted from documents in the training set, while features extracted from test set documents were used for producing prediction outputs by the CRF classifier. The Stanford Part-of-Speech Tagger [17] was used for the part of speech feature; the tagger was trained on the Wall Street Journal corpus. We did not test all possible combinations of features due to the large number of experiments required to do so; instead, we used the family of "basic" features across all tested settings (*BASIC*); we then combined "basic" features with each other feature family independently. Part-of-speech (*POS*) and regular expressions (*REG*) were considered separately and their combination with the other features was also investigated. Finally, we constructed a model that considered all combined features (*BOTH*).

For evaluating Anonym on the i2b2 dataset, the same train/test methodology used in the 2006 challenge was used in this article, with test data not analysed during the training phase. Evaluation over the CINSW and MTSamples datasets was carried out using 10-fold cross validation. Each dataset was randomly divided into 10 folds of equal size; 9 of these folds were then used to train the classifier and the remaining fold to evaluate Anonym. The process was iterated 10 times, using a different fold for testing. Evaluation measures were then averaged over the performance recorded on the testing folds.

We also evaluated Anonym across datasets, by training the CRF classifier on one dataset and testing on another. In previous work, we investigated the performance of Anonym when trained on the full i2b2 dataset (i.e., both training and testing data) and tested on a small set of pathology reports from Cancer Institute New South Wales [12]. In those initial experiments, we found that only a subset of PHIs were recognised with F-measure between 0.5 and 0.6; while other PHIs were poorly or not recognised at all. Here, we used the new CINSW dataset and the MTSamples dataset for cross-dataset effectiveness analysis.

While it seems intuitive to require high-recall performance in de-identification tasks because of the importance of removing all PHIs that may identify a person, it can be argued that high-precision is also necessary to guarantee that vital non-PHIs are not removed from the free-text documents. This is because the erroneous removal of important information such as the name of a disease (e.g., Alzheimer's disease) may render a document useless for secondary purposes. In the MTSamples dataset, strings identifying diseases or devices that may be erroneously identified as PHIs (e.g. because containing the name of a person) have been manually annotated [13]. This supports the importance of precision, beside recall, in evaluating de-identification systems. Therefore, F-measure (the harmonic mean of precision and recall) is chosen as the primary evaluation measure in the experiments of this article. To allow further comparison between our results and previous work that used the i2b2 dataset, we also report precision and recall for this dataset, along with the number of true positive (TP), false positive (FP) and false negative (FN) instances. Further analysis of the results, involving the study of precision-recall curves are left for future work.

5. Results and discussion

5.1. Performance on the i2b2 dataset

Figure 1 summaries the F-measure values recorded for all feature settings of Anonym over the testing set of the i2b2 dataset. The use of BASIC and POS features provided the best average F-measure (0.9300) on the i2b2 dataset and is represented by green dots in Figure 1. This result suggests that the pattern matching features contribute more to the de-identification effectiveness than other features, and in particular more than part-of-speech. As a reference, the best model from the 2006 i2b2 shared task [5] (the submission named Wellner 3, identified by the red dots in Figure 1), achieved an average F-measure of 0.925 on the considered PHIs, while the average F-measure obtained by the top 3 systems was 0.923. Table 4 reports a summary of the performance of Anonym on the i2b2 dataset when using the

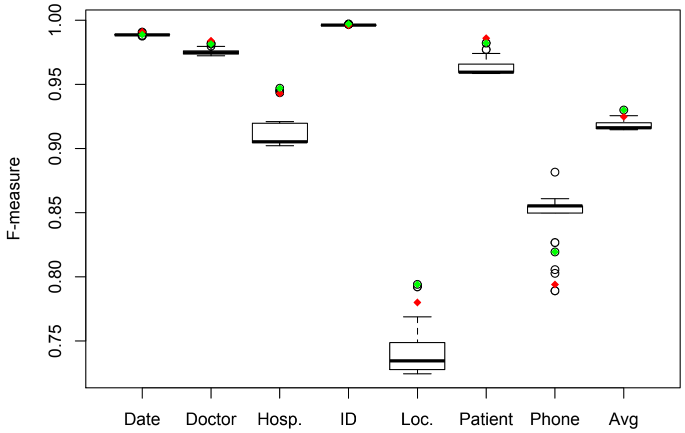

Figure 1: F-measure values obtained by Anonym using different combinations of features on the 2006 i2b2 shared task. Green points refer to the performance achieved by the best (average) combination of features (REG). Red points refer to the performance of the best system at the 2006 i2b2 challenge.

REG features.

The heights of the whiskers in the box-plots of Figure 1 represent the standard deviation in F-measure across all considered feature settings. It can be observed that PHIs for which Anonym showed higher variability with respect to feature combination are rare PHIs in the dataset (compared with Table 3). Whereas, using one combination of features in place of another is found to have little effect on those PHIs with larger number of samples (dates, doctors, IDs): these exhibit very high effectiveness and no, or marginal, variance across features. We then conjecture that Anonym can be very effective for de-identification if trained with enough samples. In addition, given that the settings that perform best overall obtained a F-measure lower than average

Table 4: Effectiveness (Precision P, Recall R, F-measure F-m, number of true positive TP, false positive FP, false negative FN) of the REG setting of Anonym on ib2b test set.

PHI Type	P	R	F-m	TP	FP	FN
DATE	.9967	.9810	.9888	2,122	7	41
DOCTOR	.9860	.9771	.9815	2,259	32	53
HOSPITAL	.9880	.9093	.9470	1,483	18	148
ID	.9983	.9958	.9971	1,195	2	5
LOCATION	.9643	.6750	.7941	162	6	78
PATIENT	.996	.9688	.9822	497	2	16
PHONE	1.0000	.6941	.8194	59	0	26
Avg.	.9899	.8859	.9300			

on the phone PHI, constructing different classifiers for identifying different PHIs may be more effective than learning a single CRF classifier.

5.2. Performance on other datasets

Table 5 summaries the effectiveness of Anonym, measured by F-measure values, on both the CINSW and MTSamples datasets across PHI types and Anonym's settings.

5.2.1. Performance on the MTSamples dataset

Anonym provides high de-identification effectiveness across all PHI types in the MTSample dataset, achieving average F-measure values up to .9807. This is obtained using the POS features or both POS and REG features; however, there is no significant difference between the effectiveness of any of the different settings of Anonym over this dataset. Phone numbers and person

Table 5: F-measure values obtained by different settings of Anonym on the CINSW and MTSamples datasets using 10-fold cross validation for training the conditional random field models. The highest F-measure obtained for each personal health identifier type within a dataset is highlighted in bold.

	CINSW dataset				MTSamples dataset			
PHI Type	BASIC	POS	REG	BOTH	BASIC	POS	REG	BOTH
AGE	.3667	.3667	.3667	.3667	.9837	.9840	**.9843**	.9840
DATE	.9400	.9362	**.9432**	.9362	.9930	**.9941**	**.9941**	**.9941**
ID	**.9911**	.9898	.9897	.9898	**.9930**	.9925	.9925	.9925
INSTIT.	.7652	**.7773**	.7670	**.7773**	**.9973**	.9959	.9970	.9959
NAME	**.9133**	.9115	.9129	.9115	.9425	**.9462**	.9443	**.9462**
PHONE	**.9699**	.9685	.9692	.9685	.9714	.9714	.9714	.9714
Avg.	.8244	**.8250**	.8248	**.8250**	.9802	**.9807**	.9806	**.9807**

names are more problematic for Anonym to de-identify in the MTSamples dataset, with F-measure performance below 0.98 for these PHI types. As reported in Table 3, in the MTSamples dataset there is only a limited number of occurrences of PHI types names and phones (286 and 7 respectively): this can explain the low performance for these PHI types. However, the PHI type ID also only occurs less than 100 times, but Anonym identifies IDs with performance above average and consistent throughout different Anonym settings. Manual analysis of ID instances in this dataset revealed that there is little variance among the format of IDs, while there is higher variance in the format of names and phones (e.g. with or without area codes). The low variance in the format of IDs produced easier patterns to be learnt by the CRF classifier. Note that in this instance, POS and REG features do not contribute higher effectiveness.

5.2.2. Performance on the CINSW dataset

The effectiveness of Anonym in de-identifying reports from the CINSW dataset is lower than that recorded on the i2b2 dataset. As in the i2b2 dataset, the highest effectiveness is achieved when only POS or both REG and POS features are used; in these settings Anonym obtains an F-measure of 0.8250. These settings, however, are found to be more effective than others only when the institution PHI type is considered: this is a relatively rare type in that dataset (177 occurrences, as reported in Table 3). The fact that part-of-speech features do sensibly improve classification for the institution type suggests that the syntactic forms used to mention pathology labs, hospitals, etc., in the CINSW dataset are very similar. Anonym has difficulty in identifying mentions of age in the CINSW dataset (F-measure of 0.3667); this is due to the small number of occurrences of ages. Higher effectiveness is obtained when more frequent PHI types are considered, e.g., IDs, names, dates. However, there is not a linear relationship between number of occurrences of a PHI and the performance of Anonym: manual inspection of the data reveals that this is because of the variance in forms between occurrences of different PHIs. For example, although in this dataset there are less occurrences of IDs than those of names, the lengths of strings associated with IDs and the strings surrounding these PHIs, i.e. context, are more similar than lengths and contexts of names. Finally, a manual inspection of the data highlighted that OCR errors do not seem to affect string associated with

PHIs, although neighbouring strings do contain OCR errors.

Table 6: F-measure values obtained by different settings of Anonym on the CINSW and MTSamples datasets when using the other dataset for training. The highest F-measure obtained for each personal health identifier type within a dataset is highlighted in bold.

	train:MTSamples; test:CINSW				train:CINSW; test:MTSamples			
PHI Type	BASIC	POS	REG	BOTH	BASIC	POS	REG	BOTH
AGE	.2128	**.2439**	**.2439**	**.2439**	0.0880	0.0800	**.1268**	**.1268**
DATE	**.5408**	.5319	**.5408**	.5319	.3688	.3694	**.3708**	.3694
ID	**.4394**	.3167	.3374	.3374	1.0000	1.0000	1.0000	1.000
INSTIT.	.5337	**.5972**	**.5972**	**.5972**	**.3011**	.2501	.2501	.2501
NAME	**.0106**	.0095	.0095	.0095	.0194	.02800	**.0306**	.0280
PHONE	.0098	.0098	.0098	.0098	**.4000**	.2222	.2222	.2222
Avg.	**.2912**	.2848	.2898	.2883	**.3629**	.3263	.3334	.3328

5.3. Performance when porting Anonym across datasets

Table 6 reports the effectiveness of Anonym when CRF models are learnt on the MTSamples datasets and tested on the CINSW dataset, and vice versa, when trained on the CINSW and tested on the MTSamples. In both sets of experiments, low de-identification effectiveness is recorded, with the best average F-measure ranging between 0.2912 and 0.3629 and the basic features providing the highest effectiveness. When testing on the CINSW dataset, Anonym achieves the highest effectiveness on dates and institutions. While the part-of-speech and regular expression features are overall outperformed by the use of basic features only, these do provide consistently better performance on the institution PHI type. This result is similar to that reported in Table 5: these features were also found to provide higher de-identification effectiveness for mentions of institutions. Very poor performance is obtained when de-identifying names and phone numbers on the CINSW dataset; this is likely due to the small amount of training occurrences for these PHI types present in MTSamples. Higher average effectiveness is recorded when testing on MTSamples and training on CINSW, where ages are infrequent in the training dataset. This explains the low performance for this type of PHI, although a small gain is obtained by regular expression features over the BASIC and POS features for this PHI. Surprisingly, all IDs in the MTSamples dataset are recognised with perfect precision (F-measure 1.0). A manual analysis of IDs in the training and testing dataset, however, did not unveil strong similarities between the formats of IDs, or their contexts.

6. Summary of findings

In answer to our first research question (RQ1, Section 4), the experiments and analysis revealed that Anonym is effective across datasets, with overall F-measures ranging between 0.81 and 0.98. Anonym does provide state-of-the-art effectiveness on the i2b2 2006 shared task, with an average F-measure value higher than the best system at i2b2.

We found that the effectiveness of Anonym is not solely dependent on the amount of available training occurrences, as F-measures across PHIs did not seem to be linearly correlated with the number of occurrences (RQ2). A manual analysis of the data suggested that in most cases it is the combination of training size and variation among PHIs format that influences effectiveness.

Cases have been highlighted in our experiments where Anonym has achieved higher de-identification effectiveness over less frequent but more cohesive PHI types than over those PHI types that were more frequent.

No one feature set consistently performed best in our experiments (RQ3). While the features produced by regular expressions provided the highest effectiveness on the i2b2 dataset, a similar finding was not confirmed on the CINSW and MTSamples data, where part-of-speech features (as well as both POS and REG) were found to be more effective. When combining these two datasets for training and testing, the basic set of features provided instead the highest effectiveness.

In terms of robustness (RQ4), the de-identification capabilities of Anonym are found to adapt to different data (e.g., discharge summaries vs. pathology reports), conventions (United States vs. Australia), and quality (typed vs. OCRed). However, for Anonym to be effective, enough training data has to be provided and this has to be consistent with the documents that are expected to be given for de-identification. In absence of enough training samples or when data comes from a different dataset, Anonym achieves unsatisfactory performance.

7. Conclusions

Accessing and sharing EHRs is fundamental for fostering data mining, information retrieval and natural language processing research that aims to improve health service delivery and medical knowledge discovery. These possibilities are however hindered by the presence of personal health information in free-text health records; de-identification of this information is required for the secondary use of this data for research. Manual de-identification is time and resource consuming. In this article, we have evaluated a software solution for the de-identification of EHRs, called Anonym. We have evaluated Anonym across datasets and settings, and found that Anonym is effective and robust to variation in the data if trained with enough and representative samples.

Acknowledgements

The authors wish to acknowledge Danielle Mowery and Wendy W. Chapman for providing access to the MTSamples dataset.

References

[1] Demner-Fushman D, Chapman W, McDonald C. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics 2009;42(5):760–1.

[2] Prokosch H, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods of Information in Medicine 2009;48(1):38–44.

[3] O'Keefe C, Connolly C. Privacy and the use of health data for research. Medical Journal of Australia 2010;193(9):537–41.

[4] Dorr D, Phillips W, Phansalkar S, Sims S, Hurdle J. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods of Information in Medicine 2006;45(3):246–52.

[5] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association 2007;14(5):550–63.

[6] Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. Journal of the American Medical Informatics Association 2007;14(5):564–73.

[7] Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine 2008;42:13–35.

[8] Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010;10(1):70–1.

[9] Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Journal of the American Medical Informatics Association 2013;20(1):77–83.

[10] Hanauer D, Aberdeen J, Bayer S, Wellner B, Clark C, Zheng K, et al. Bootstrapping a de-identification system for narrative patient records: Cost-performance tradeoffs. International Journal of Medical Informatics 2013;82:821–31.

[11] Boström H, Dalianis H. De-identifying health records by means of active learning. In: Elhadad N, Hauskrecht M, editors. ICML Workshop

on Machine Learning for Clinical Data Analysis. Edinburgh, Scotland (UK); 2012, p. 1–4.

[12] Zuccon G, Strachan M, Nguyen A, Bergheim A, Grayson N. Automatic De-Identification of Electronic Health Records: An Australian Perspective. In: Suominen H, editor. The 4th International Workshop on Health Document Text Mining and Information Analysis. Sydney, NSW, Australia; 2013, p. 1–6.

[13] South B, Mowery D, Ferrandez O, Shen S, Suo Y, Zhang M, et al. On the Road Towards Developing a Publicly Available Corpus of De-identified Clinical Texts. In: Hersh W, editor. American Medical Informatics Association Annual Symposium Proceedings. Chicago, IL (USA); 2012, p. 1950–1.

[14] Nadeau D, Sekine S. A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes 2007;30(1):3–26.

[15] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Brodley C, Pohoreckyj Danyluk A, editors. Proceedings of the 18th International Conference on Machine Learning. Williamstown, MA (USA); 2001, p. 282–9.

[16] Collins M. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In: Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting on Association for Computational Linguis-

tics. Philadelphia, PA (USA): Association for Computational Linguistics; 2002, p. 489–96.

[17] Toutanova K, Manning C. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Schütze H, Su KY, editors. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. Hong Kong: Association for Computational Linguistics; 2000, p. 63–70.

[18] Zuccon G, Nguyen A, Bergheim A, Wickman S, Grayson N. The impact of OCR accuracy on automated cancer classification of pathology reports. In: Maeder A, Martin-Sanchez F, editors. Studies in Health Technology and Informatics; vol. 178. 2012, p. 250–8.